Enhancing Modality-Agnostic Representations via Meta-learning for Brain Tumor Segmentation

Aishik Konwer¹, Xiaoling Hu¹, Joseph Bae², Xuan Xu¹, Chao Chen², Prateek Prasanna²
¹Department of Computer Science, Stony Brook University
²Department of Biomedical Informatics, Stony Brook University
{akonwer,xiaolhu,xuaxu}@cs.stonybrook.edu
{joseph.bae, chao.chen.1, prateek.prasanna}@stonybrook.edu

Abstract

In medical vision, different imaging modalities provide complementary information. However, in practice, not all modalities may be available during inference or even training. Previous approaches, e.g., knowledge distillation or image synthesis, often assume the availability of full modalities for all subjects during training; this is unrealistic and impractical due to the variability in data collection across sites. We propose a novel approach to learn enhanced modality-agnostic representations by employing a meta-learning strategy in training, even when only limited full modality samples are available. Meta-learning enhances partial modality representations to full modality representations by meta-training on partial modality data and meta-testing on limited full modality samples. Additionally, we co-supervise this feature enrichment by introducing an auxiliary adversarial learning branch. More specifically, a missing modality detector is used as a discriminator to mimic the full modality setting. Our segmentation framework significantly outperforms state-of-the-art brain tumor segmentation techniques in missing modality scenarios.

1. Introduction

Multiple medical imaging modalities/protocols are required to provide complementary diagnostic cues to clinicians. For instance, multiple Magnetic Resonance Imaging (MRI) sequences (henceforth referred to as modalities), namely native T1, post-contrast T1 (T1c), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR) are used together to understand the underlying spatial complexity of brain tumors and their surroundings [3, 5]. Deep learning approaches [21, 36, 10, 58, 51, 49] have found great success in multimodal brain tumor segmentation and treatment response assessment. These conventional brain tumor segmentation methods perform well only when all four acquisition modalities are available as input (i.e., in the full modality setting). However, in clinical practice, often only a subset of modalities are available due to issues including image degradation, motion artifacts [18], erroneous acquisition settings, and brief scan times. Hence it is crucial to develop robust modality-agnostic methods which can achieve state-of-the-art performance in missing modality settings, i.e., when different modalities are unavailable during inference or even training.

Recently, a plethora of works has been proposed to address missing modality scenarios for brain tumor segmentation. Two major categories include: 1) Knowledge distillation: These methods [43, 23, 52, 50, 2] learn privileged information from a teacher network trained on full modality data $D_f$ for all subjects or simulate partial modality data $D_m$ from $D_f$. On the contrary, our approach works in a limited full modality setting, i.e., $|D_f| \leq |D_m|$.

![Figure 1: Comparison of the paradigms generally adopted by existing missing modality approaches (left) vs. ours (right) for brain tumor segmentation. $N$ and $n$ refer to the number of subjects (patients) with partial and full modalities, respectively. Previous methods either utilize full modality data $D_f$ for all subjects or simulate partial modality data $D_m$ from $D_f$. On the contrary, our approach works in a limited full modality setting, i.e., $|D_f| \leq |D_m|$.

Figure 1: Comparison of the paradigms generally adopted by existing missing modality approaches (left) vs. ours (right) for brain tumor segmentation. $N$ and $n$ refer to the number of subjects (patients) with partial and full modalities, respectively. Previous methods either utilize full modality data $D_f$ for all subjects or simulate partial modality data $D_m$ from $D_f$. On the contrary, our approach works in a limited full modality setting, i.e., $|D_f| \leq |D_m|$.
generator. This can be very unrealistic; in real-world applications, most studies only have very limited full modality data, far from sufficient for training. In this paper, we focus on a more realistic setting: most training data is only partial modality data, i.e., having a few modalities missing. We ask the following question: How do we efficiently learn from a large amount of partial modality data and a small amount of full modality data (see Fig. 1b)?

Another category recently rising in popularity is Shared Latent Space modeling [22, 14, 28, 7, 13, 56, 57, 53, 20, 30, 60, 8]. These methods learn a shared latent representation from partial modality data. However, the quality of the learned representation can be limited by the heterogeneity of available modalities. The learned representation will be biased towards the most frequently available modalities and essentially overlook minority modalities (i.e., modalities that appear less frequently in training). This will inevitably lead to sub-optimal performance on test data, especially with minority modalities. To compensate for this undesirable bias, these methods often resort to segmenting all modalities individually from the shared representation, ultimately requiring full modality for all cases during training.

These observations, further summarized in Tab. 1, motivate us to design a modality-agnostic method that can fully utilize partial modality data. Through the usage of the meta-learning strategy, our method learns enriched shared representations that are generalizable and not biased towards more frequent modalities, even with limited full modality data.

Our core idea is based on the meta-learning technique [16]. Meta-learning provides an effective framework to learn to perform multiple tasks in a mutually beneficial manner. We consider segmentation with each partial modality input combination as a different task, yielding \(2^M - 1\) meta-tasks for \(M\) modalities. By learning all meta-tasks in parallel, meta-learning ensures the network generates modality-agnostic representations. Thanks to meta-learning, tasks depending on rare modalities can be significantly improved even with limited training data. This maximally mitigates the bias against rare modalities. Meanwhile, we propose using a small amount of full modality data only during meta-testing. Meta-testing is introduced as an intermediate step in meta-learning to boost the generalization performance of the model across different tasks. Using full modality data, albeit limited, in meta-test can maximally leverage such data to enhance the representation quality of the model. This innovative meta-learning design ensures we learn with a large amount of partial modality data and only a small amount of full modality data, with negligible partial modality bias.

Recently, a meta-learning approach [31] performed classification with missing modalities. They predict the prior weights of modalities via a feature reconstruction network, the quality of which is indirectly dependent on the number of full modality samples. This method is unsuitable for our segmentation framework since conventional approaches (PCA [40], K-Means [33]) cannot be used to cluster the priors in a high-dimensional latent space. Moreover, [31] deals with only two input modalities and considers them individually as meta-tasks, while we construct a heterogeneous task distribution with different combinations of inputs respecting the heterogeneity settings of real-world data.

We also employ a novel adversarial learning technique that further enhances the quality of the generated shared latent space representation. Previous GAN-based approaches [42] reconstruct the missing modalities in image space; this leads to the impractical requirement of full modality as ground truth for training. Our task is achieved in latent space by designing the discriminator as a multi-label classifier. The discriminator predicts the presence/absence of modalities from the fused latent representation performing a binary classification for each modality. Our ultimate goal is to hallucinate the full modality representation from the hetero-modal feature space. Note that due to the hetero-modal nature of the data, the number of available modalities can vary dramatically across subjects. To address this, we utilized a channel-attention weighted

**Table 1: Advantages of our approach over existing frameworks.** We are able to train in a limited full modality (FM) setting (with \(\leq 50\%\) FM samples), and learn an unbiased mapping that is unaffected by the proportion of any given modality in training.

<table>
<thead>
<tr>
<th>Category</th>
<th>Can handle limited FM?</th>
<th>Learns Unbiased mapping?</th>
</tr>
</thead>
<tbody>
<tr>
<td>KD [23, 52, 50, 2]</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>GAN [42, 54, 59, 25, 55]</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Shared (others) [8, 13, 56]</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>SMIL [31]</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Shared (Ours)</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Figure 2: Framework overview.** \(D_m\) (partial modality) and \(D_{\text{full}}\) (full modality) are used as inputs for encoder-decoder networks in the meta-train and meta-test phase, respectively. Partial modality representations are adapted to the full modality domain via: 1) meta-optimization of gradients in both data, and 2) adversarial learning based on predictions by a modality absence classifier.
fusion module that can accept a varying number of representations as input but generates a single fused output. Overall, our contributions can be summarized as follows:

- We propose a meta-learning paradigm to train with hybrid data (partial and full modalities) and also enhance the learned partial modality representations to mimic a full modality representation. This is accomplished by meta-training on partial modality data while fine-tuning on limited full modality data during the meta-test. Such a training strategy overcomes the over-reliance on full modality data, as well as succeeds in learning an unbiased representation for all missing situations.

- We introduce a novel adversarial learning strategy to further enrich the shared representations in the latent space. It differs from other generative approaches that synthesize missing images and demand full modality ground truths for training. Our approach does not necessitate reconstructing missing modality images.

2. Related Work

Segmentation with missing modalities. Incomplete data is a long-standing issue in computer vision; it particularly has significant implications in medical vision. Due to privacy concerns and budget constraints, one or more modalities (audio/visual/text) [27, 15, 11, 48, 6] of a given sample may not be available. In this work, we focus on partial medical imaging modalities for brain tumor segmentation. Existing methods on complete multi-modal brain tumor segmentation [21, 36, 10, 58, 51, 49] perform poorly in realistic hetero-modal settings.

Researchers have broadly used three techniques to perform brain tumor segmentation from missing modalities including knowledge distillation (KD) [43, 23, 52, 50, 2], generative modeling [42, 54, 59, 25, 55] and shared representation learning [22, 14, 28, 7, 13, 56, 57, 53, 20, 30, 60, 8]. ACN [52] trains a separate teacher-student pipeline for each subset of modalities. Among the generative models, MM-GAN [42] uses a U-Net to impute missing modalities while a PatchGAN learns to discriminate between real and synthesized inputs. A major drawback of KD and GAN-based approaches is their inability to perform well when all modalities are not present for a subject during training. Moreover, the unstable and non-converging nature of a 3D generator may lead to degraded quality of synthesized images, eventually affecting downstream performance. Our method belongs to the third category, i.e., shared latent space models. In [22, 14], the authors compute variational statistics to construct unified representations for segmentation. Multi-source information is modeled using a correlation constraint [60] or region-aware fusion blocks [13] to encode shared representations. Recent frameworks [57, 56] in this genre advocate for exploiting intra/inter-modality relations through graph and transformer-based modeling. Such approaches usually lack flexibility for adaptation to all missing scenarios. They yield sub-optimal performance due to failure in the retrieval of discriminative features generally existing in full modality data. Furthermore, these approaches can learn biased mappings among the available modalities leading to poor generalizability for modalities not encountered in training.

Meta-learning. Meta-learning algorithms [16, 37, 47] are inspired by human perception of new tasks. Optimization-based meta-learning techniques [16, 37, 1] have gained popularity since they can easily encode prior information through an optimization process. Model-agnostic meta-learning (MAML) [45] is the most commonly used algorithm under this category, due to its flexible application to any network trained through gradient descent. Researchers have widely adopted MAML frameworks to generalize a model to new tasks, unseen domains, and enriching input features in multimodal scenarios [29, 32]. SMIL [31] introduces a Bayesian MAML framework that attains comparable classification performance across both partial and full modality data. However, their approach requires prior reconstruction of the missing modalities. HetMAML [9] can handle heterogeneous task distributions, i.e. different modality combinations for input space but fails to attain generalizable performance across partial and full modalities. Inspired by the above two approaches, we propose a modality-agnostic architecture that can not only accept hetero-modal inputs but also enhance their representations with the additional information present in a full set of modalities. This leads to better segmentation performance for any hetero-modal input instance.

Domain adaptation. Domain adaptation refers to the training of a neural network to jointly generate both discriminative and domain-invariant features in order to model different source and target data distributions [17, 4, 19, 46, 26]. Authors in [17] leverage an auxiliary domain classifier to address the domain shift. Inspired by this approach, we design our discriminator as a modality absence predictor. Similar to Sharma et al. [42], we feed our discriminator with the correct modality code as ground truth, while the generator is provided an ‘all-one’ full modality code impersonating all missing scenarios. They yield sub-optimal performance in missing modality situations. This results in enhanced representations that boost downstream performance in missing modality situations.

3. Methodology

Overview. Given heterogeneous modalities as input, our goal is to build a modality-agnostic framework that can be robust to missing modality scenarios, and achieve performances comparable to a full modality setting. We have lim-
Distribution model becomes more robust to missing scenarios at inten
tient during training. This training paradigm ensures the
tion of subjects, some modalities are dropped for each pa-
nario where full modalities may only be available for a frac-
putes for a patient. To simulate a real-world clinical sce-
3.1. Meta-Learning for Feature Adaptation

Our proposed approach is shown in Fig. 3. Meta-learning has been shown to be an efficient compu-
tational paradigm in dealing with heterogeneous training
data [9], or conducting feature adaptation between differ-
ent domains [31]. To this end, we adapt model-agnostic meta-learning to leverage information from both partial and full modality data. This strategy is elaborated in Sec. 3.1. Because we need to generate a common fused representation for each hetero-modal input combination, our architecture incorporates a simple and elegant feature aggregation module (see Sec. 3.3).

3.1. Meta-Learning for Feature Adaptation

Suppose we have a total of $M$ MRI modalities as in-
puts for a patient. To simulate a real-world clinical sce-
nario where full modalities may only be available for a frac-
tion of subjects, some modalities are dropped for each pa-
itent during training. This training paradigm ensures the
model becomes more robust to missing scenarios at in-
fERENCE. Thus we construct a heterogeneous task distribution $P(T)$ that is a collection of $k$ task distributions: $P(T^1), P(T^2), \ldots, P(T^k)$. Each such distribution $P(T^i)$ has a distinct input feature space related to a specific subset of modalities. We however exclude the full modality subset from the task distribution due to its utilization in meta-
testing, as explained in the following paragraph. Overall, $k$ types of task instances can be sampled from $P(T)$, where $k = 2^M - 2$.

Algorithm 1 Modality-Agnostic Meta-Learning

1: **Input**: Training dataset $D$ is divided into two cohorts of subjects with partial/missing and full modalities re-
spectively $D = \{D^m, D^f\}$; $\beta$ is the learning rate.
2: **Initialise**: Initialise $\theta_g, \phi_d = \{\theta_{dis}, \theta_{dec}\}, \alpha$
3: **Output**: Optimized meta-parameters $\{\theta_g, \phi_d, \alpha\}$
4: while not converged do
5:   Sample a batch of tasks $T^t \sim \{P(T)\}$
6:   for each task $T^t$ do
7:     Evaluate inner loop loss: $L_{\text{miss}}(\theta_g, \phi_d; \mathcal{D}_m^i)$
8:     Adapt: $\theta_g^* = \theta_g - \alpha \nabla_{\theta_g} L_{\text{miss}}(\theta_g, \phi_d; \mathcal{D}_m^i)$
9:     Compute outer loop loss: $\mathcal{L}^{\text{ull}}(\theta_g^*, \phi_d; \mathcal{D}_f^i)$
10:   end for
11: Update meta-parameters: $(\theta_g, \phi_d, \alpha) \leftarrow (\theta_g, \phi_d, \alpha) - \beta \nabla_{(\theta_g, \phi_d, \alpha)} \sum_{T^t} \mathcal{L}^{\text{ull}}(\theta_g^*, \phi_d; \mathcal{D}_f^i)$
12: end while

Formally, we have a hetero-modal training dataset $D$ which we divide into two cohorts of subjects $\{D^m, D^f\}$ containing partial and full modalities, respectively. The goal is to effectively learn from both types of data. We con-
struct a batch of subjects $\mathcal{D}_m^i$ corresponding to each $P(T^i)$. The pair of subjects and their corresponding task remains fixed over all epochs. Only the modalities which are not included in a task get dropped for that particular subject.

Shared encoders are used along with a fusion module
to produce a modality-agnostic representation. In our case, both encoder and fusion modules jointly constitute the generator $E_{d}$ (parameterized by $\theta_{d}$). An MLP-based classifier network, parameterized by $\theta_{d}$, is employed as a discriminator as explained in Sec. 3.2. For clarity, parameters of the discriminator and decoder network, $\{\theta_{d}, \theta_{dec}\}$, are collectively symbolized as $\phi_{d}$. Our aim is to obtain an optimal generator parameter $\theta_{g}$ through task-wise training on $D^{m}_{i}$ by reducing the inner loop objective $L^{\text{miss}}_{g_{i}}$.

$$\theta_{g}^{*} = \theta_{g} - \alpha \nabla_{\theta_{g}} L^{\text{miss}}_{g_{i}}(\theta_{g}, \phi_{d}; D^{m}_{i}),$$

(1)

where $\alpha$ is a learnable rate for inner-level optimization. The optimized model is expected to perform better on $D^{f}$. The goal of the updated framework is to accomplish the outer loop objective $L^{\text{full}}$ across all sampled tasks:

$$\min_{\theta_{g}, \phi_{d}} \sum_{T_{i}} L^{\text{full}}(\theta_{g}^{*}, \phi_{d}; D^{f}).$$

(2)

Both the inner and outer loop losses are kept the same, referring to the generator and discriminator losses, $L_{E}$ and $L_{dis}$. By forcing the partial model trained to perform well on full modality data, we implicitly target the recovery of relevant information for better segmentation in missing modality scenarios. This partial to full modality mapping in feature space, is further strengthened by the introduction of a domain-adaptation inspired feature enrichment module (Details in Sec. 3.2). All three meta parameters $(\theta_{g}, \phi_{d}, \alpha)$ are henceforth meta-updated by averaging gradients of outer loop loss over a meta-batch of tasks:

$$(\theta_{g}, \phi_{d}, \alpha) \leftarrow (\theta_{g}, \phi_{d}, \alpha) - \beta \nabla_{(\theta_{g}, \phi_{d}, \alpha)} \sum_{T_{i}} L^{\text{full}}(\theta_{g}^{*}, \phi_{d}; D^{f}).$$

(3)

Thus during meta-training, the model tunes its initialization parameter to achieve improved generalizability across all missing modality tasks. During meta-test, by finetuning with full modality data, we map the learned feature representations to the full modality space. Different from MAML, the pretrained model is directly evaluated on datasets where subjects contain a fixed subset of modalities (one of the tasks $T^{i}$ already encountered in meta-training) at inference. The training process is summarized in Alg. 1.

### 3.2. Adversarial Feature Enrichment

Considering that full modality data contains richer information, we enforce encoder outputs to mimic full representations, irrespective of the limited input combination. The modality encoders and fusion module can be collectively considered as a shared generator $E$. We introduce an MLP-based multi-label classifier as our discriminator $D$.

The objective of $D$ is to predict the absence/presence of modalities from the fused embedding $F_{\text{fused}}^{R}$ at the bottleneck level. $D$ utilizes Binary-Cross-Entropy loss $L_{\text{BCE}}$, and sigmoid activation to output $M$ binary predictions $\hat{d}$, denoting whether a modality is available or not. While calculating the discriminator loss $L_{\text{dis}}$ indicated below, the ground truth variable $T_{\text{real}}$ is a vector of size $M$ which reflects the true combination of modalities available at input for that iteration. For example, assuming that $M = 4$, and only first two modalities are available, $T_{\text{real}} = \{1, 1, 0, 0\}$.

$$L_{\text{dis}} = \sum_{z=1}^{D^{m}+D^{f}} L_{\text{BCE}}(\hat{d}_{z}, T_{\text{real}}).$$

(4)

The generator loss is a combination of segmentation loss and an adversarial loss used to train the generator to fool the discriminator. We consider a dummy ground truth variable $T_{\text{dummy}}$. In order to encourage the generator to encode representations that confuse or “fool” the discriminator into inferring that all modalities are present, we set $T_{\text{dummy}} = \{1,1,1,1\}$, masquerading all generated representations as full modality representations. Thus $D$ pushes the generator $E$ to agnostically produce full modality representations.

$$L_{E} = \lambda_{1} L_{\text{seg}} + \lambda_{2} \sum_{z=1}^{D^{m}+D^{f}} L_{\text{BCE}}(\hat{d}_{z}, T_{\text{dummy}}).$$

(5)

### 3.3. Modality-Agnostic Feature Aggregation

We aim to utilize multiple modalities (which vary in number per patient) and derive a common fused representation. Individual encoders $E_{1}, E_{2}, ..., E_{n}$ having shared parameters are trained to extract features from each of the $n$ available patient-specific modalities, where $1 \leq n \leq M$. These features $F_{1}^{l}, F_{2}^{l}, ..., F_{n}^{l}$ obtained from the corresponding levels ($l$) of each encoder are passed into a feature aggregation module.

![Feature Aggregation Module](image)

Figure 4: Illustration of the feature aggregation module. Modality $F_{2}^{l}$ is missing. $F_{1}^{l}$ and $F_{3}^{l}$ are passed through global average pooling (GAP) operation and eventually fed into an MLP to generate the shared representation $F_{\text{fused}}^{l}$. The individual encoded representations undergo a global average pooling (GAP) operation and eventually fed into an MLP to generate the shared representation $F_{\text{fused}}^{l}$.
concatenated to form a $M$-dimensional vector $\gamma$. This is achieved by imputing zeros in the channel information of $(M - n)$ missing modalities. $\gamma$ is mapped to the channel weights of $M$ modality features through a multi-layer perceptron (MLP) and sigmoid activation function, $\sigma$. These modality-specific weights multiplied with the corresponding features give rise to the aggregated representation, $F_{\text{fused}}$ (in Fig. 4), which is eventually used as input to the decoder for segmentation. Our aggregation module exploits the correlation among available modality representations to create a unified feature that best describes the tumor characteristics of a subject. Detailed explanations with equations regarding this module can be found in the supplementary (Sec. 12).

We adopt a Swin-UNETR [49] architecture that employs soft Dice loss [35] to perform voxel-wise semantic segmentation. The segmentation loss function $L_{\text{seg}}$ is defined as follows:

$$L_{\text{seg}}(G, P) = 1 - \frac{2 \sum_{v=1}^{V} \sum_{u=1}^{U} G_{u,v}^{2} + \sum_{u=1}^{U} P_{u,v}^{2}}{V}$$

where $V$ is the number of classes and $U$ is the number of voxels. $F_{u,v}$ and $G_{u,v}$ refer to the predicted output and one-hot encoded ground truth for class $v$ at voxel $u$, respectively.

4. Experiment Design and Results

To validate our framework in various missing scenarios, we evaluate brain tumor segmentation results on all fifteen combinations of the four image modalities for a fixed test set. The average score is also reported for comparisons.

Datasets. We use three segmentation datasets from BRATS2018, BRATS2019, and BRATS2020 challenges [34]. They comprise 285, 335, and 369 training cases respectively. All subjects have $M = 4$ MR sequences. We perform 3D volumetric segmentation with images of size $155 \times 240 \times 240$. Pre-processing details are contained in the supplementary (Sec. 15). The segmentation classes include whole tumor (WT), tumor core (TC), and enhancing tumor (ET). Additional segmentation results on two non-BRATS hetero-modal cohorts (with different medical imaging modalities such as CT and MRI) are also reported in the supplementary (Sec. 16).

Implementation details. The experiments are implemented in Pytorch 1.7 [39] with three 48 GB Nvidia Quadro RTX 8000 GPUs. We drop modalities during training to construct the missing-modality dataset $D^m$. We randomly sample a set of subjects from $D^m$ assigned to each task distribution $P(T)$. For a given subject, only those modalities which are not present in its associated task distribution are dropped. We ablate the fraction of subjects reserved for the full modality dataset $D^f$ (Fig. 7a). Our method adopts a modified Swin-UNETR [49] housing up to 4 encoders $E_1, E_2, E_3, E_4$ which are Swin transformers (see supplementary Sec. 12). Both MLPs for the discriminator and feature aggregation are fully connected networks whose hidden layer dimensions are 48 and 64 respectively. The images are first resized to $128 \times 128 \times 128$, which is kept consistent across all compared methods. Features are extracted from 5 different levels of each encoder. For training and testing cohorts, we randomly split BRATS2018 into 200 and 85 subjects, BRATS2019 into 250 and 85 subjects, and BRATS2020 into 269 and 100 subjects, respectively. The batch size per task is kept as 1. During meta-training, we consider a metabatch size of 8, i.e., our meta-batch comprises 8 different modality combinations, each representing a separate task $T_i$. More details can be found in the supplementary (Sec. 15). During inference, the meta-pretrained model is evaluated on test sets where all subjects have a fixed subset of modalities. The discriminator is discarded at inference.

Performance metrics. Dice similarity coefficient (DSC $\uparrow$) (Tab. 2) and Hausdorff Distance (HD95 ↓) (see supplementary Sec. 9) are used to evaluate segmentation performance.

4.1. Comparisons with State-of-the-art

Quantitative results: In Tab. 2, we compare our approach with SOTA methods including HeMIS [22], U-HVED [14], D2-Net [53], ACN [52], RFNet [13] and mmFormer [56] on BRATS2018. HeMIS, U-HVED, and D2-Net learn a biased mapping among available modalities and hence perform poorly compared to ACN which co-trains with the full modality of all samples. Recent shared latent space models like mmFormer and RFNet perform comparably to ACN. They either focus on learning inter-modal correlations or tumor region-aware fused representations. Our method utilizes full modality data from only 50% samples, and yet outperforms these approaches. We thus excel in efficient utilization of full modality data. In comparison with the second-best approach in WT, TC, and ET, our average DSC shows improvements of 0.89% (over mmFormer), 1.96% (over ACN), and 1.68% (over ACN), respectively. Although ACN pursues a KD-driven approach to achieve the partial-to-full modality mapping, it ends up building a combinatorial number of models dedicated to each subset. This leads to a highly ineffective solution which is also based on the impractical scenario that all training samples contain full modality data. In our framework, we mimic this distillation learning even in a shared latent model through efficient application of meta-learning and adversarial training. It can be seen from Tab. 2 that our method surpasses all other approaches in 39 out of 45 multi-modal combinations across the three tumor regions despite being trained with only 50% full modality samples. Our results are statistically significantly better (t-test, $p < 0.05$) than HeMIS [22], U-HVED [14], D2-Net [53]. Other methods (RFNet [13],
Effectiveness of adversarial and meta-learning. We perform several ablations to evaluate and justify the contribution of each proposed module in our architecture. First, we consider different combinations of available modalities on BRATS2018. Our approach trains with 50% full modality samples while others use different settings, identical to ours (Tab. 3). It can be observed that our method outperforms the second-best approach by 11.78%, 12.93%, and 9.72% in DSC on the WT, TC, and ET regions, respectively, on BRATS2018, clearly achieving SOTA performance. Evaluation via HD95 metric on BRATS2018 can be found in the supplementary (Sec. 9). Comparison with three SOTA methods on BRATS2020 are presented in Tab. 5; Compared to the second-best approach, our method outperforms the second-best approach by 4.52%±26.52%, 4.07%, respectively. Modalities present are denoted by •, the missing ones by ◦. Statistically significant results with p-values \( \leq 0.05 \) are denoted by *.

### 4.2. Ablation Studies

**Qualitative results:** In Fig. 5 we visualize the segmentation masks predicted by U-HVED, RFNet, and ours from mmFormer [56], ACN [52]) require full modality input for all samples, whereas ours does not. However, for a fair comparison, we further trained [13, 56, 52] in a 50% full modality setting, identical to ours (Tab. 3). It can be observed that our method outperforms the second-best approach by 11.78%, 12.93%, and 9.72% in DSC on the WT, TC, and ET regions, respectively, on BRATS2018, clearly achieving SOTA performance. Evaluation via HD95 metric on BRATS2018 can be found in the supplementary (Sec. 9). Comparison with three SOTA methods on BRATS2020 are presented in Tab. 5; Compared to the second-best approach, the average DSC of the three tumor areas is boosted by 11.78%, 12.93%, and 9.72% in DSC on the WT, TC, and ET regions, respectively, on BRATS2018, clearly achieving SOTA performance. Evaluation via HD95 metric on BRATS2018 can be found in the supplementary (Sec. 9).

**Effectiveness of adversarial and meta-learning.** We perform several ablations to evaluate and justify the contribution of each proposed module in our architecture. First, we consider different combinations of available modalities on BRATS2018. Our approach trains with 50% full modality samples while others use different settings, identical to ours (Tab. 3). It can be observed that our method outperforms the second-best approach by 11.78%, 12.93%, and 9.72% in DSC on the WT, TC, and ET regions, respectively, on BRATS2018, clearly achieving SOTA performance. Evaluation via HD95 metric on BRATS2018 can be found in the supplementary (Sec. 9). Comparison with three SOTA methods on BRATS2020 are presented in Tab. 5; Compared to the second-best approach, the average DSC of the three tumor areas is boosted by 11.78%, 12.93%, and 9.72% in DSC on the WT, TC, and ET regions, respectively, on BRATS2018, clearly achieving SOTA performance. Evaluation via HD95 metric on BRATS2018 can be found in the supplementary (Sec. 9).
remove both the Adversarial and the Meta-training strategies to perform segmentation from only fusion of available modalities. We thus formulate a baseline, mDrop, where we include our feature-aggregation block to generate a fused representation from available modalities. mDrop solely learns the intra-model relations through transformer encoders and inter-modal dependencies through channel-weighted fusion. The average DSC of our model outperforms mDrop by 5.45%, 6.71%, and 10.47% in the three tumor regions (Tab. 4). Hence it is evident that solely the modality-agnostic representations obtained from fusion of available modalities cannot generate accurate segmentations. This necessitates feature enrichment to improve the quality of the fused representation. We develop two variants through gradual introduction of our discriminator and meta-learning strategies as enrichment techniques. Both variants surpass mDrop considerably (Tab. 4). Meta-learning (MetaL) proved to be better since we built the heterogeneous task distribution with modality combinations (to reduce bias) and also explicitly adapted to the full modality feature space efficiently. Finally, we arrive at an end-to-end meta-learning framework that also benefits from auxiliary supervision provided by the adversarial discriminator.

**Evaluation of enhanced representations.** To evaluate the quality of enhanced representations, we designed a simple experiment. We first extracted the bottleneck fused representations $F_{\text{fused}}^B$ of 50 test subjects for both scenarios of full modality (where all are present) and partial modality (where only T1c, T2 are present). This was done for both mDrop as well as our approach. The fused representations were fed into a classifier trained to predict the probabilities of a modality’s presence. The average probabilities obtained from our method attain comparable distributions across partial and full modality settings (Fig. 6), depicting the desired enhancement of $F_{\text{fused}}^B$. However, for mDrop, probabilities of T1c and T2 being present are much higher than T1, and FLAIR in the missing scenario. Hence the relevant information from the latter two modalities is being lost. The red boxes depict how our probability for predicting T1 is considerably higher than mDrop even in T1-missing scenario. Due to the meta-learning strategy incorporated while training on hybrid data, we hypothesize that our network is robust to the ratio of full modality samples used in training. We compare against

**Robustness to full modality setting.** Due to the meta-learning strategy incorporated while training on hybrid data, we hypothesize that our network is robust to the ratio of full modality samples used in training. We compare against

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average DSC (%)</th>
<th>p-value (10^{-*})</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACN[52]</td>
<td>WT 65.81, 0.011* 75.66, 0.011* 47.36, 1.63*</td>
<td>mDrop 81.67 72.41 52.06</td>
</tr>
<tr>
<td>RFNet[13]</td>
<td>WT 73.96, 0.011* 62.37, 0.011* 50.24, 0.80*</td>
<td>RFNet[13] + GAN 84.49 75.38 55.64</td>
</tr>
<tr>
<td>mmFormer[56]</td>
<td>WT 75.34, 0.011* 66.19, 0.011* 52.81, 7.64</td>
<td>mmFormer + GAN 85.96 77.23 59.85</td>
</tr>
<tr>
<td>Ours</td>
<td>WT 87.12 79.12 62.53</td>
<td>Ours + GAN 87.12 79.12 62.53</td>
</tr>
</tbody>
</table>

Table 3: Comparison (DSC%, p-value) on BRATS2018 with 50% full modality

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average DSC (%)</th>
<th>p-value (10^{-*})</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACN[52]</td>
<td>WT 65.81, 0.011* 75.66, 0.011* 47.36, 1.63*</td>
<td>mDrop 81.67 72.41 52.06</td>
</tr>
<tr>
<td>RFNet[13]</td>
<td>WT 73.96, 0.011* 62.37, 0.011* 50.24, 0.80*</td>
<td>RFNet[13] + GAN 84.49 75.38 55.64</td>
</tr>
<tr>
<td>mmFormer[56]</td>
<td>WT 75.34, 0.011* 66.19, 0.011* 52.81, 7.64</td>
<td>mmFormer + GAN 85.96 77.23 59.85</td>
</tr>
<tr>
<td>Ours</td>
<td>WT 87.12 79.12 62.53</td>
<td>Ours + GAN 87.12 79.12 62.53</td>
</tr>
</tbody>
</table>

Table 4: Ablation study demonstrating effectiveness of major components.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average DSC (%)</th>
<th>p-value (10^{-*})</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACN[52]</td>
<td>WT 65.81, 0.011* 75.66, 0.011* 47.36, 1.63*</td>
<td>mDrop 81.67 72.41 52.06</td>
</tr>
<tr>
<td>RFNet[13]</td>
<td>WT 73.96, 0.011* 62.37, 0.011* 50.24, 0.80*</td>
<td>RFNet[13] + GAN 84.49 75.38 55.64</td>
</tr>
<tr>
<td>mmFormer[56]</td>
<td>WT 75.34, 0.011* 66.19, 0.011* 52.81, 7.64</td>
<td>mmFormer + GAN 85.96 77.23 59.85</td>
</tr>
<tr>
<td>Ours</td>
<td>WT 87.12 79.12 62.53</td>
<td>Ours + GAN 87.12 79.12 62.53</td>
</tr>
</tbody>
</table>

Table 5: Comparison (DSC%, p-value) on BRATS2020.

Figure 6: Comparison of baseline and enhanced features ACN, RFNet, and mmFormer by varying the full modality count from 100% to 40% (Fig. 7a). In order to retain sufficient samples for each combination task in meta-training, we assume that at least 50% of the subjects have partial modalities. Hence we show our results only on 50% and 40% proportions of full modality data. The fact that even with 50% full modality samples, we match the evaluation scores of SOTA at 100% setting is noteworthy. A sharp degradation can be noticed in the average WT DSC of SOTA once the number of full modality data decreases. On the other hand, our method shows only a minor drop of 0.29%. This is due to ACN being heavily dependent on the full modality for knowledge distillation. RFNet and mmFormer require full modality data as input to the network. They under-fit since their overall sample count decreases. Our method efficiently utilizes even limited samples of full modality data for feature adaptation in meta-testing. Owing to the above reasons, our approach is resilient to change in full modality proportion. Results for other tumor regions are provided in supplementary (Sec. 8).

Figure 7: Ablation studies.

(a) Ablation results for varying % of full modality in training. (b) Ablation results for varying % of FLAIR in training.

**Bias to presence of a specific modality.** Our model is robust to the scenario when a modality appears rarely during training. Tab. 6 demonstrates that when only 35% of FLAIR is considered for training, our method consistently outper-
Table 6: Results for rare occurrence of FLAIR in training.

Table: [Data table]

<table>
<thead>
<tr>
<th>M</th>
<th>FLAIR</th>
<th>T1</th>
<th>T1c</th>
<th>T2</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>U-HVED</td>
<td>76.61</td>
<td>78.94</td>
<td>83.98</td>
<td>80.47</td>
</tr>
<tr>
<td></td>
<td>D2-Net</td>
<td>82.44</td>
<td>83.79</td>
<td>85.16</td>
<td>85.55</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>86.53</td>
<td>88.77</td>
<td>90.05</td>
<td>89.41</td>
</tr>
<tr>
<td>TC</td>
<td>U-HVED</td>
<td>50.23</td>
<td>55.46</td>
<td>57.35</td>
<td>71.83</td>
</tr>
<tr>
<td></td>
<td>D2-Net</td>
<td>47.19</td>
<td>59.98</td>
<td>59.65</td>
<td>81.23</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>67.26</td>
<td>72.28</td>
<td>72.64</td>
<td>84.03</td>
</tr>
<tr>
<td>ET</td>
<td>U-HVED</td>
<td>17.58</td>
<td>20.39</td>
<td>32.17</td>
<td>67.75</td>
</tr>
<tr>
<td></td>
<td>D2-Net</td>
<td>9.37</td>
<td>11.78</td>
<td>14.09</td>
<td>68.14</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>42.58</td>
<td>44.27</td>
<td>47.14</td>
<td>76.12</td>
</tr>
</tbody>
</table>

5. Conclusion

We present a novel training strategy to address the problem of missing modalities in brain tumor segmentation under limited full modality supervision. We adopt meta-learning and formulate modality combinations as separate meta-tasks to mitigate the bias towards modalities rarely encountered in training. We distill discriminative features from full modality data in the meta-testing phase, thereby discarding the impractical omnipresence of full modalities for all samples. This mapping is further co-supervised by novel adversarial learning in latent space, that guarantees the generation of superior modality-agnostic representations. In the future we will validate our method on other downstream tasks such as radiogenomics classification [44] and treatment response prediction [41].

6. Acknowledgements

Reported research was partially supported by NIH R21CA258493-01A1 and NSF CCF-2144901. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

encoder-decoder for joint modality completion and segmentation. In *MICCAI*, 2019. 2, 3, 6, 7, 8


[38] Ozan Oktay, Jo Schlemper, Loic Folgoc, Matthew Lee, Matthias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Hammerla, Bernhard Kainz, Ben Glover, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas. In *MIDL*, 2018. 9


[40] Karl Pearson. Li.ii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 1901. 2


[56] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In MICCAI, 2022.